DATA WAREHOUSING

Peter J. Weyman
VMARK SOFTWARE INC.

The Case for a Process-Driven Approach to Data Warehousing

Is your company in the midst of a business information crisis? I use the word "crisis" to describe any system in which the demand for service exceeds that system's capacity to provide service. Many companies' information technology (IT) groups were thrown into a state of crisis when users of business intelligence tools or other decision support software tried to use the "all that information we've been spending all that IT money on." In most cases, attempts to connect non-IT users with raw IT systems have been less than successful.

The business information crisis is the latest in a series of roughly decade-long crises that have plagued IT departments and fueled the explosive growth of the high-tech industry. In the fifties, the "business automation crisis" resulted from the desire to eliminate manual labour from routine business processes such as cost accounting and payroll management. In response to this crisis, a number of mainframe software application packages were written that automated these business processes.

Each of these applications managed its own data files, so their emergence resulted in the "data centralization crisis" of the sixties. The issues that defined this crisis were the data redundancy, inconsistency, security, integrity and sharing problems that resulted from having each application manage its own data. The technology industry responded with database management systems, which centralized data management.

Early database management systems, such as DATACOM-DB, IMS and IMS were essentially very complicated programming systems, and their proliferation helped bring on the "software crisis" of the seventies. There simply weren't enough talented programmers to meet the demand. This crisis motivated a number of programmer productivity advancements, including database products based on the relational data model. The relative simplicity of this data model was supposed to make applications easier to develop, and data sharing among applications easier.

Relational database management systems delivered on the promise of easier application development, but their poor performance precipitated the "OLTP performance crisis" of the eighties. The industry responded with Entity-Relationship (E-R) data modelling, which resulted in tremendous performance gains. E-R modelling drives all redundancy out of data, and divides it into many discrete entities, each of which becomes a table. The result is that transactions are highly localized in their effect, minimizing data access steps that negatively impact performance. Another result is that virtually all business meaning is normalized out of the data as well This renders data as stored in OLTP systems practically useless for business analysis purposes. That brings us to the current business information crisis, which is that IT organizations have been focused on automating business processes, not on building a business information infrastructure.

The Data Warehouse Response

The industry's response to the business information crisis has been the data warehouse, a specialized database that provides integrated, relevant, consistent information about the business, modelled for use by decision support and other business analysis software products.

The chances are, that if you are building or considering building a data warehouse, you are doing so as part of a strategic IT initiative to improve the flow of business information to users of decision support software within your company. As implied above, users of these products require *access* to *integrated* organizational data that is *consistent*, and *organized* for their use. The challenge is to meet these requirements, starting with "dirty" and inconsistent data from multiple "un-integrated", even incompatible, data sources that have been optimized into obscurity. All without impacting production system performance! Figure 1 illustrates the extent of this challenge

The Business Information Infrastructure

As can be seen in Figure 1, implementing a successful data warehouse is more than buying some decision support software and a data warehouse database. These products rely on a framework for gathering data, organizing it into meaningful information and delivering it for evaluation in the context of the business. This framework is called a business information infrastructure (BII), and is sometimes referred to as the data warehousing *process*.

Something else to keep in mind is that production systems are rarely designed to keep track of historical information. They are designed to automate processes "in the now", and can only provide current operational information. Answers to many strategic and tactical business questions require more than just access to current information, however, they also require access to comparable historical information over varying periods in time. For example, "How do sales of pink widgets in the eastern region compare to year ago sales?" Gathering and keeping track of historical data is an important BII function.

With this background, data warehousing can be seen as a process with several essential stages: *modelling* data for the data warehouse, *extracting* data from the source databases, *cleansing* data to filter out or repair defective records, *transforming* data from the production model into the warehouse model, *integrating* data with data from other sources, and *loading* the data into the data warehouse database. With the exception of data model-

►

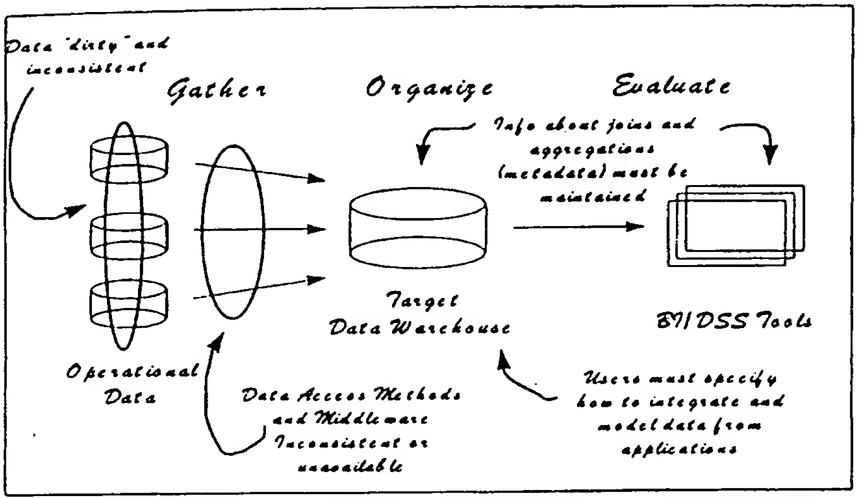
Data "dirty" and inconsistent
Gather
Organize
Evaluate
Info about joins and aggregations (metadata) must be maintained
Target Data Warehouse
BI/DSS Tools
Operational Data
Data Access Methods and Middleware Inconsistent or unavailable
Users must specify how to integrate and model data from applications

Figure 1. Effective decision support means gathering relevant data, organizing it into meaningful information, and evaluating the information in the context of the business. Extracting, cleansing, transforming and integrating data, while keeping track of information that describes the data are tasks that make or break data warehouses. The figure shows common problems that face DSS implementers.

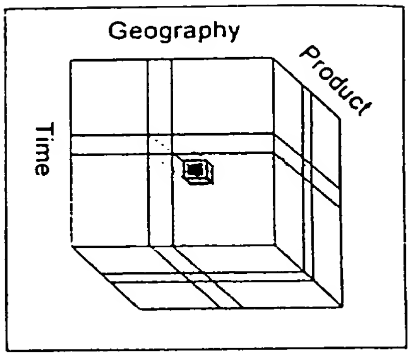

Figure 2. The dimensional model of data describing a business. Each intersection point stores the fact or measurement for a particular combination of the dimensions. The answer to the question "How many pink widgets did the eastern region sell during the third quarter?" can be found at the intersection of (product, geography, time) corresponding to (pink widgets, eastern region, third quarter).

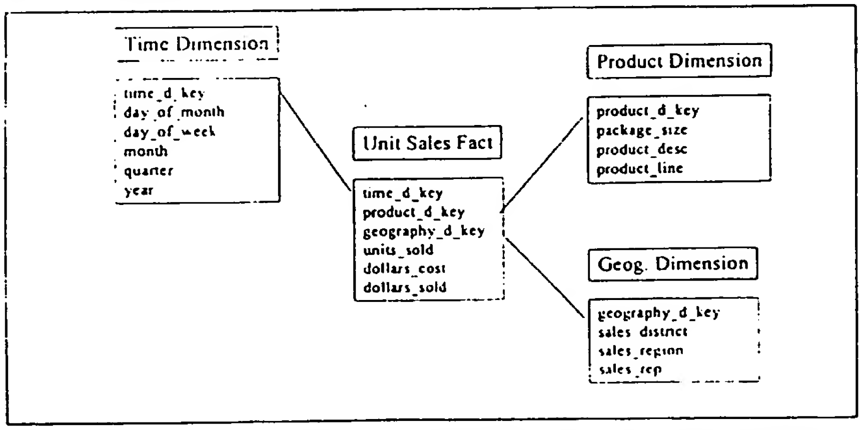

Figure 3. The Star Join Schema. The primary key of each dimension table is a foreign key in the unit sales fact table. Each dimension table stores descriptions of the measurement dimensions of the business. Each descriptive field in a measurement dimension can be used to constrain queries and provide row headers for the user's response. The fact table stores the numerical measurements.

ling, these are activities that must occur on an on-going basis, usually daily. Data modelling is a relatively infrequent activity, but will occur when changes to the warehouse data model or supporting BII are required.

Dimensional Data Models

Most data warehouses are implemented with *dimensional data models.* To understand this model, let's look into how we might store data to respond to queries like "How many pink widgets did the eastern region sell during the third quarter?" This query tells us that the analyst thinks of the business in three dimensions: product (pink widgets), geography (eastern region) and time (third quarter). A dimensional representation of facts to answer queries like this is shown in Figure 2.

This model differs radically from the fully-normalized entity-relation models used by OLTP systems. The dimensional model is often called the star join schema, because the diagram used to represent the model looks like a star, as shown in Figure 3.

Dimensional models are ideal for data warehouses because they are simple for users of business intelligence tools and other decision support software to understand. Unfortunately, they can be difficult to design. To state the obvious, success at this step requires frequent, high-quality interaction with the users of the warehouse. Understanding how they model the business (that is, the kind of questions they ask, the kind of metrics they care about) is key to providing a usable data model. Remember too that the model is likely to change, as the business evolves, or as competitors or market trends elevate the importance of a previously ignored metric. The model has the additional advantage of being extensible — adding dimensions and attributes to dimensions will rarely "break" existing software and queries.

Data Exraction and Cleansing

This is the process of getting the data needed for the warehouse from operational databases, archives and external data sources. In practical systems, a variety of extraction techniques are required: live queries on operational or mirrored data, reading database table dump tapes, interpreting transaction logs, and sometimes, writing custom programs to extract data from legacy application files.

It's important to note that operational databases need to be off-line for production purposes during extraction, to avoid data inconsistencies. The on-line high-availability nature of operational databases means that this stage must be accomplished on a tight schedule during the database "back up and maintenance window", when the database is scheduled to be off-line for production use.

Because of the large size of many operational databases, and the relatively short periods of time available for data extraction, little transformation of the data occurs during this process. Efficiency of this process can be greatly enhanced if only changed data is extracted. That is, only data that has changed since the last extraction is extracted. This is known as *changed data capture.* If space permits, a copy of the data as of the last extraction can be kept on the source machine, and a program written to compare current data with the stored image. Alternatively, replication techniques can be used to store changed records for later extraction to the data warehouse.

Businesses that purchase commercial data, such as competitive performance information, or market surveys, or who want to use government supplied data, will need to extract data from whatever format the external source prescribes.

Operational data is notoriously dirty. Negative inventory values, mis-spelled names and missing fields are common. Data not required by the operational data system (for example, as a key value) is much more likely to be dirty, since it is unlikely to have been validated on input (to avoid performance penalties associated with validation). Data cleansing is the process by which invalid records are filtered out or repaired on their way to the data warehouse. At this relatively late time in the life of a data record, it is very difficult to repair records with missing fields. For this reason, one of the most effective preventative steps that can be taken is to make these fields required on input. Required analysis fields should be considered when revisions to data entry procedures are developed as part of business process re-engineering efforts.

An interesting, if subtle source of data "dirty-ness" is organizational change over time. Remember that the data warehouse builds up a historical view of the business. This can be difficult when re-organization changes geographical districts and department names, businesses are acquired and divested, product codes are re-used or changed and so on. As was the case above, any data not required by the operational system may not be validated on input (e.g. names on different bank accounts), and so may not be comparable. The extraction and cleansing process must either deal with the data automatically, or through human intervention.

Data Transformation and Integration

As has become obvious by now, transforming the data from fully normalized, dis-integrated operational schemata into a dimensional schema for users of business analysis tools is essential to the data warehousing process. This process is made more difficult when data comes from different data sources (applications, databases or systems), because data for comparison as part of the integration and transformation process is less likely to be stored or even named in a comparable fashion.

Coding, decoding, adding descriptive information obviated and previously only available to E-R CASE tools, adjusting representations (for example, adding or removing separator characters in account numbers) all need to be done. A complicating factor is that data from different sources may be stored at different levels of granularity. For example, unit sales in one data source may be stored by day, while in another, they may be stored by week. Rationalization examples abound: day vs. week vs. month, zip code vs. county vs. region, demographic by age vs. occupation vs. income and so on.

Sometimes data for analysis must be synthesized from data that has been stored for operational purposes. For example, operational data stores often provide *level* information (e.g., quantity in stock), when analysis requires *event* information (e.g. units shipped or received). This information can be synthesized by comparing current operational data with previous snapshots.

Another consideration is the speed of query execution for users of the business intelligence front end tool. If frequent queries aggregate (roll up) the data along certain dimensions (e.g. into quarters, from weeks), it may be worthwile to pre-compute and store some of these aggregations during the data transformation process. Usage statistics gathered during use of analysis tools may suggest additional pre-aggregations to accelerate common analysis queries.

This is the step in the data warehousing process where data from different data sources is integrated into the dimensional analysis schema. Challenges to this process (often a multi-table join operation) include differing key values from different data sources, differing data coding formats (a customer ID in a sales application may be coded differently from the same customer ID in a customer service database), differing views of data granularity (time series constraints) and so on. Because this operation depends heavily on data comparison, it is particularly sensitive to dirty data.

Data from data sources that cannot be queried on-line (e.g. archived legacy data files on tape) poses an additional challenge in that the transformation process must itself build a view of the data that can be effectively transformed and integrated into the dimensional analysis schema.

In addition to these logical constraints, the transformation process is time-bound as well. As mentioned above, data extraction must occur while the source database is off-line for production use. This limits the time window for extraction and transformation processing. In larger systems, extraction and transformation must occur at the rate of hundreds or thousands of records per second!

Data Loading

Like operational databases, data warehouse databases must be available for use for most of the day. The warehouse data must present a

▶

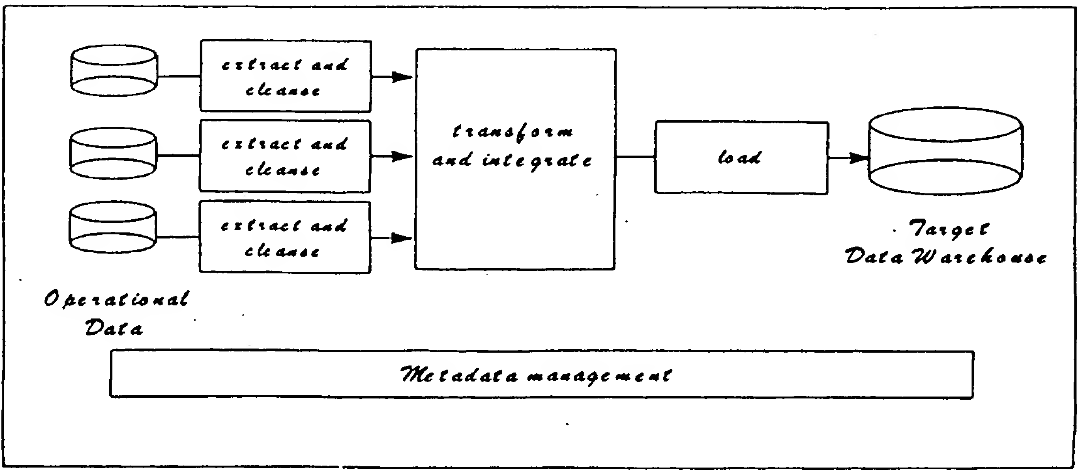

Figure 4. The data warehousing process. Operational data is extracted and cleansed, integrated with data from other data sources and transformed into a form suitable for analysis, then loaded into the target data warehouse. Throughout the process, metadata which describes the data in source, intermediate and target forms, as well as the transformation and integration steps must be maintained and accessible for process, analysis tool and audit trail use.

DATA WAREHOUSING – SURVEY

consistent and static view of the business: analysis and research involves long queries. and multiple queries. and the validity of conclusions depends on underlying data not changing. For this reason, loading the data warehouse must take place while the warehouse database is off-line for analysis use. During the loading process for a dimensional analysis database. new facts are added to fact tables, old facts may need to be removed or archived due to size limits, and newly pre-computed aggregations need to replace previously calculated values. Further, the speed of analysis queries depends critically on fast indexing and query optimization using the indices.

Data loading can be either a "push" type activity. in which a loading process INSERTs. UPDATEs, and DELETEs records in the target data warehouse. or a "pull" activity. in which the data warehouse database itself pulls data through the loading process, often as a table load operation.

DataStage, a Solution for Data Warebouse Development and Maintenance

By now, it is apparent that implementing a data warehouse is more than just buying a data warehouse database engine and a business analysis front end. Successful data warehouse implementation requires implementation of a robust and repeatable process, as shown in Figure 4. VMARK's DataStage is a product intended to address this need, (see Page 00). ■

SMSmail for Psion

Dynamical Systems Research Limited has launched SMSmail, a Psion 3a/3c compatible software package that enables Psion users to send and receive Internet e-mail, messages and files to and from a palm-top computer using SMS.

SMSmail uses 'SMS Link Cable' that connects the Nokia 2110, 8110 or other compatible phones to a Psion.

SMS stands for Short Message Service, a text messaging technology that is in place on Vodafone, Cellnet and Orange digital networks, and on other GSM and PCN networks world-wide. SMS allows the transfer of text messages of up to 160 characters in length from one mobile phone to another.

SMSMail costs £50.

DSR Ltd. Tel: +44 171 584 0084. Fax: +44 171 584 5442

Jargon confuses even IT Pros.

A Benchmark Research survey of Times 1000 companies indicates that IT companies have succeeded in confusing IT experts as well as workers.

The survey, sponsored by software companies Cognos and VMARK. was based on 400 separate telephone contacts with both end users and IT professionals in the UK's largest companies. Its findings prove that confusion over computer terminology affects those who make their living from IT even more than those who have to use it in their jobs.

Whilst 400 organisations agreed to participate in the research, 300 were not able to complete the survey due to total lack of knowledge about the technology subject. Of those that could participate, Benchmark found that only 29% of non-IT respondents could correctly define a 'Data warehouse'. The figure for IT specialists was worse still. just 24% correctly defined this key technology.

The research. carried out in October. consisted of qualitative face-to-face depth interviews followed by twenty minute Computer Aided Telephone Interviews (CATI). Those interviewed were asked what their level of knowledge about 'data warehousing' was.

Only six percent of IT professionals claimed extensive knowledge of the subject, compared to four percent of end users, yet 34% of those surveyed said their organisations already had a data warehouse. All respondents were given six definitions of the term Data warehousing. In total, 73% got the answer wrong. One senior IT professional even referred to the software as hardware.

The top three issues among 100 respondents to a telephone questionnaire showed marked differences between those with first-hand knowledge and those without.

The prime concerns of those without data warehousing were higher cost of implementation (69%), data quality (61%) and data security (56%). Companies using a data warehouse cited lack of end-user training (59%) ahead of higher cost of implementation (53%), data quality (50%) and an inability to meet end-user expectations (50%) as their biggest challenges.

Dr. Greg Bohlen, research manager at Benchmark Research: "These apparently contradictory findings are, in fact, linked. The very reason higher cost is perceived as a potential barrier to using data warehouses is the inability of many UK IT specialist to cost items such as end-user training."

Other factors metnioned included the risk of using unproved technology, ranked fifth by those who did not yet have a data warehouse (sixth by those who have) controlling access to data, mis-use of data and a lack of knowledge about data warehousing among the IT department. Overall a larger share of those with data warehouses had no particular concerns about data warehousing (18%) than those without data warehouse (11%).

IT management has little understanding of some of the fundamentals required for success prior to the start of the project. because, historically. IT departments have been responsible for implementing systems which have helped manage the day-to-day operations of the business. OLTP systems were designed to process orders. ship product. raise invoices, etc., etc., but were never designed to produce data which could be used for Data warehouse purposes. Managing the complexities of moving from one environment to another is one of the keys to achieving success with Data warehousing.

Cognos have announced the publication of a new book aimed at businessmen and women who wanted to see practical benefits from data warehousing. Called 'Twenty four ways to impact your bottom line in 90 days', the book gives practical examples of how data warehouse technology can save companies money.

Copies of a slide report represented by Dr. Greg Bohlen of Benchmark Research. and the 'Twenty Four Ways' book are available on request.

Cognos Limited. Tel: 01344 486668. Fax: 01344 485124

Data Warehousing in Retailing

The challenge facing many retailers today is not collecting information, but in using it to improve the sales and profitability of their business. according to a new report on the impact of data warehousing by retail consultants. Management Horizons Europe.

Management Horizons' report lists the top 10 challenges which retailers face as they move into the year 2000. Winning and maintaining a competitive edge by constantly improving customer satisfaction will remain as the primary focus for retailers.

According to the report, although retailers are responding to these challenges by implementing an array of customer-driven and cost reduction programmes, in many cases they have not seen a return because of a lack of effective technology. The job of a data warehouse is to untangle the maze of operational. financial. customer and external marketplace data.

'Data Warehousing: Information on Demand' is available free of charge from *Management Horizons Europe, Waverley House, Lower Square, Isleworth. Tel: 0181 560 9393. Fax: 0181 568 6900.* ■

XP-002091745

.03.1995
446-447 2

Data Extraction and Transformation for the Data Warehouse

A Presentation by
Cass Squire, Channel Director of Professional Services
Prism Solutions, Inc.

Corporations worldwide are finding that understanding and managing rapidly growing, enterprise-wide data is critical for making timely decisions and responding to changing business conditions. To manage and use business information competitively, many companies are establishing decision support systems built around a data warehouse of subject-oriented, integrated, historical information.

In order to understand why the data warehouse must replace old legacy applications for effective information processing, it is necessary to understand the root causes of the difficulty in getting information in the first place. The first difficulty in getting information from the base of old applications is that those old applications were shaped around business requirements that were relevant as much as twenty-five years ago. These applications that were shaped yesterday do not reflect today's business.

The second reason why older applications are so hard to use as a basis for information is that those applications were shaped around the clerical needs of the corporation. A clerically focused application of necessity does not have the historical foundation required to support a long-term view.

Another reason why the clerical perspective of applications does not support management's need for information is that the clerical community focuses on detailed data. While detailed data is fine for the day-to-day clerical needs of the organization, management needs to see summary data in order to identify trends, challenges and opportunities.

Yet another reason why the clerical perspective o applications does not suffice for management's need fo information is that the clerically-oriented applications were built an application at a time, and there was little or nc integration from one application to the next. The result is that the old legacy applications cannot easily or reliably be combined to produce a unified perspective of data.

For these basic reasons, the older foundation of applications will not suffice as a basis for the important informational processing that organizations need to do in order to become efficient, competitive corporations. Nothing short of an entire change in architecture and a fundamental restructuring of the applications foundation will suffice.

Fortunately there is an alternative architecture, which consists of a separation of processing into two broad categories—operational processing and decision-support processing. At the heart of decision-support (DSS) processing is the structure known as the data warehouse.

The data warehouse contains data which has been gathered and integrated from the legacy systems environment. There are different levels of data within the data warehouse. Some data is very detailed. Other data is summarized. Other older detailed data is placed in secondary storage. In addition there is a component of the data warehouse known as "meta data." Meta data, or information about data, is a directory as to what the contents of the data warehouse are and where the contents came from.

Six major steps are involved in implementing a data warehouse: 1) building the data warehouse data model, 2) defining the system of record, or "best data" for the warehouse, 3) designing the physical data warehouse, 4) creating the transformation programs, 5) loading and maintaining the data warehouse, and 6) building and maintaining directory of meta data.

Building a data warehouse requires extraction of data from legacy systems, operational applications and external sources. As data passes from the system of record into the data warehouse, it passes through a set of programs called integration and transformation programs. These programs take application-oriented data and turn the data into corporate data. The integration and transformation programs perform a wide variety of functions, such as—

- reformatting data,
- recalculating data,
- modifying key structures of data,
- adding an element of time to data warehouse data,
- identifying default values of data,
- supplying logic to choose between multiple sources of data,
- summarizing data,
- tallying data,
- merging data from multiple sources, and so forth.

Among the challenges involved in data extraction and transformation include the fact that source data exists in heterogeneous mainframe and UNIX-based environments. The navigational paths of these legacy systems and operational applications are complex. What's more, inconsistencies between naming conventions, business rules and definitions used must be resolved. In addition, data integrity and quality must be verified and maintained throughout the transformation process.

The integration and transformation programs are very susceptible to maintenance as they need to be modified each time the operational environment changes or each time the data warehouse environment itself changes.

There are several advantages to automating the development and maintenance processes. Automated tools can reduce implementation time and cost substantially by eliminating the need for manual programming. Structured code generation increases productivity, promotes consistency across programs and allows quick response to change. Data integrity is maintained by performing data extraction and transformation automatically rather than manually. Finally, automated tools actively capture and maintain meta data related to source files, output tables, transformations and mappings, providing a record of changes and enhancements to the data warehouse over time.

One reason why meta data management is mandatory for the data warehouse environment is due to the size of the warehouse. Meta data serves as the "card catalog," helping users navigate the data warehouse and find relevant information for analysis. Another reason why meta data is so important is because of the horizon of time that is managed in the data warehouse. It is typical for 5 to 10 years of data to be stored in the data warehouse. One of the implications of managing such a lengthy time period of data is that the structure of data will change over time. Meta data stores the context of this historical information.

Meta data exists at two levels in the data warehouse—at the developer level and at the end user level. Developer meta data is used by the developer to manage and control the development and maintenance process. End user meta data resides on the data warehouse platform itself and is available to the end user as a regular part of the access and analysis of the warehouse.

SIGMOD ' 95,San Jose , CA USA
© 1995 ACM 0-89791-731-6/95/0005..$3.50